Google Cloud

Gen AI: Beyond the Chatbot Review

Congratulations on completing the first course of the AI Hypercomputer earning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

AI Hypercomputer is a supercomputing system that is optimized to support your artificial intelligence (AI) and machine learning (ML) workloads. It's an integrated system of performance-optimized hardware, open software, ML frameworks, and flexible consumption models.

Create
Generate new content.

Summarize
Condense information into concise summaries

Discover
Find information at the right time

Automate
Automate previously manual tasks

Foundation models

Large AI models trained on massive datasets, allowing them to be adapted to many tasks. They are the basis of gen AI.

Key Features of Foundation Models:

- Trained on diverse data.
- Flexible to a wide range of use cases.
- Adaptable to specialized domains through additional, targeted training.

Prompting

Prompting is the method of interacting with and guiding foundation models by providing them with instructions or inputs to generate desired outputs.

Ask Gemini ➤

Gemini is a Google gen AI model that powers many different solutions.

- Gemini app
- Google Workspace with Gemini
- Gemini for Google Cloud

Vertex AI is Google Cloud's unified ML platform. It empowers you to build, train, and deploy ML applications. Vertex AI gives you access to generative AI models (such as Gemini) and lets you tune them to meet your needs, and then deploy them.

Google is an AI first company

- Gen AI tools are integrated across Google's ecosystem.
- Google ensures you stay updated with the latest AI advancements.
- Google provides an-ecosystem that puts security and ethics at the forefront.
- Build on Google Cloud's enterprise-grade foundation.
- Google's open approach gives you flexibility and choice in your AI solutions.

Gen AI strategy: Combine a top-down approach (leadership setting the vision and strategy) with a bottom-up approach (employees identifying practical applications and providing feedback).

- **Strategic focus:** Prioritize focused gen AI implementations with clear business value.
- **Exploration:** Encourage experimentation and collaboration to discover valuable gen AI applications.
- **Responsible AI:** Establish ethical guidelines and ensure secure and responsible AI development.
- **Resourcing:** Invest in data strategy, leverage existing resources, and develop AI talent.
- **Impact:** Measure gen AI's impact on business goals and demonstrate tangible benefits.
- **Continuous improvement:** Continuously refine gen AI solutions based on feedback and data.

Google Cloud

Gen AI: Unlock Foundational Concepts

Congratulations on completing the second course of the Gen AI Leader learning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

Artificial intelligence (AI): Building machines that can perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making.

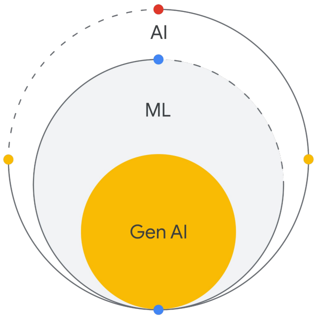
Machine learning (ML): A subfield of AI where machines learn from data to perform specific tasks.

Generative AI: An application of ML that focuses on creating new content.

Deep learning: A subset of ML that uses artificial neural networks with many layers to extract complex patterns from data.

Foundation models: Powerful ML models trained on massive amounts of unlabeled data, allowing them to develop a broad understanding of the world.

Large language models (LLMs): A type of foundation model that is designed to understand and generate human language.

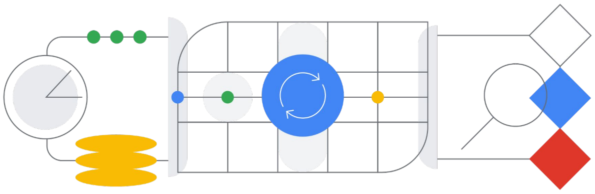


Data is information that can come in many forms: numbers, dates, text descriptions, and even images or sounds.

- **Structured data:** Data that is organized and easy to search, often stored in relational databases.
- **Unstructured data:** Data that lacks a predefined structure and requires sophisticated analysis techniques.
- **Data quality:** Ensure your data is accurate, complete, consistent, and relevant.
- **Data accessibility:** Data for model training needs to be readily available, usable, and in the proper format.

ML lifecycle

- **Data ingestion and preparation:** The process of collecting, cleaning, and transforming raw data into a usable format for analysis or model training.
- **Model training:** The process of creating your ML model using data.
- **Model deployment:** The process of making a trained model available for use.
- **Model management:** The process of managing and maintaining your models over time.



ML has **three primary learning approaches**:

- **Supervised learning** trains models on labeled data to predict outputs for new inputs.
- **Unsupervised learning** uses unlabeled data to find natural groupings and patterns.
- **Reinforcement learning** learns through interaction and feedback to maximize rewards and minimize penalties.

Responsible AI

- **Secure AI:** Protecting your AI applications from harm.
- **Ethical AI:** Ensuring your AI applications don’t cause harm and are used in an ethical manner.

The **Secure AI Framework (SAIF)** helps organizations manage AI/ML model risks and ensure security.

Google Cloud

Gen AI: Navigate the Landscape

Congratulations on completing the third course of the Gen AI Leader learning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

Before starting your gen AI project consider:

Needs:

- **Scale**: How many users will there be?
- **Customization**: How specialized is this AI?
- **User interaction**: How will users engage?
- **Privacy**: How sensitive is the data?
- **Latency**: What response time can you have?
- **Connectivity**: What is your connectivity?

Resources:

- **People**: Do you have access to AI expertise?
- **Money**: What's your budget?
- **Time**: What are your project timelines?

Platform: The foundation for building and scaling AI initiatives.

Vertex AI: Google Cloud's unified machine learning (ML) platform designed to streamline the entire ML workflow. It provides the infrastructure, tools, and pre-trained models you need to build, deploy, and manage your ML and generative AI solutions.

Infrastructure: The foundation upon which everything else rests. It provides the core computing resources needed for generative AI. This includes the physical hardware (like servers, GPUs, and TPUs), along with the essential software needed to train, store, and run AI models.

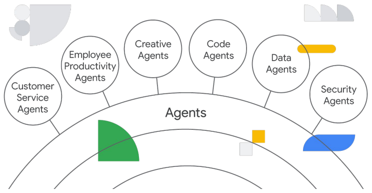
AI on the edge: You can run AI solutions on infrastructure (devices or servers) closer to where the action is happening.

Google provides tools like **Lite Runtime (LiteRT)** to help developers deploy AI models on edge devices.

Gemini Nano is Google's most efficient and compact AI model, specifically designed to run on devices.

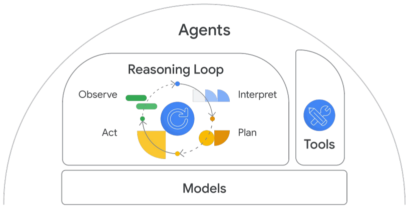
Gen AI landscape

- Gen-AI-powered applications
- Agent
- Platform
- Model
- Infrastructure



Agent: A gen AI agent is an application that tries to achieve a **goal** by **observing** the world and **acting** upon it using the tools it has at its disposal.

It does this using:



- **Reasoning loop:** An iterative process where the agent observes, interprets, reasons, and acts, often using prompt engineering
- **Tools**: Functionalities that allow the agent to interact with its environment, such as accessing and processing data or interacting with software or hardware.
- **Model:** The brains of the AI system, which consist of various algorithms that learn patterns from data and can make predictions or generate new content.

Vertex AI gives you options for how to handle AI models for your project.

- **Model Garden:** Pick from existing Google, third-party, or open-source models.
- **Model Builder:** Train and use your own models. Go fully custom and create and train models at scale using an ML framework. Or use AutoML to create and train models with minimal technical knowledge and effort.

Google Cloud

Gen AI Apps: Transform Your Work

Congratulations on completing the fourth course of the Gen AI Leader learning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

Prompting techniques

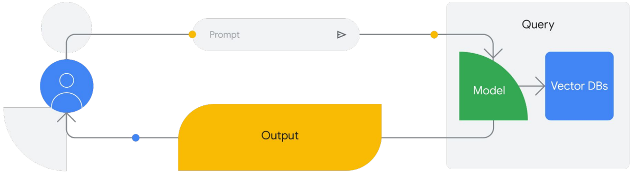
- **Zero-shot:** Asking the model to complete a task with no prior examples.
- **One-shot:** Providing the model with one example to learn from.
- **Few-shot:** Giving the model multiple examples to learn from.
- **Role:** Assigning a persona to the model to influence its style, tone, and focus.
- **Prompt Chaining:** Engaging in a back and forth conversation with the AI.

Model guidance and refinement

Grounding: Connecting the AI's output to verifiable sources of information.

RAG: Retrieval-Augmented Generation

1. **Retrieving relevant information:** The AI model retrieves relevant information from a vast knowledge base.
2. **Generating output:** The model then uses this retrieved information to generate the final output.



Gemini tooling for personal productivity

The Gemini app is Google's generative AI chatbot, where you can get help with writing, planning, learning, and more.

Gemini for Google Workspace integrates gen AI into familiar Workspace apps, allowing you to do things like compose emails in Gmail, generate images in Slides, and summarize notes in Meet.

Gemini for Google Cloud is your AI assistant for Google Cloud. It can help you write and debug code, manage and optimize cloud applications, analyze data in BigQuery, and strengthen your security posture.

NotebookLM allows you to upload your files and then acts as a research assistant, summarizing key points, answering questions, and generating ideas, all while staying grounded in your source material.

Streamlining Prompting Workflows

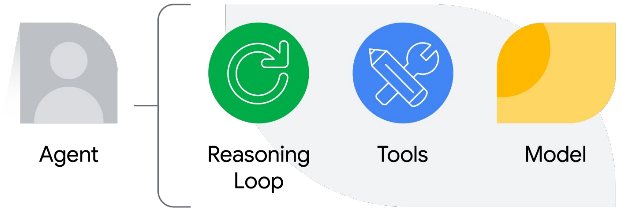
- **Reusing Prompts**: Saving prompts as templates for repeated use.
- Leveraging **Prompt Chaining**: Continuing conversations within the same chatbot to maintain context.
- Using **Saved Info** in Gemini: Storing specific information for the model to use consistently.
- **Gems** are personalized AI assistants within Gemini. They provide personalized responses tailored to specific instructions, streamline workflows like templates, prompts, and guided interactions.

Google Cloud

Gen AI Agents: Transform Your Organization

Congratulations on completing the fifth course of the Gen AI Leader learning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

Agent components



Types of agents

- **Deterministic (traditional):** Agents that are built with predefined paths and actions.
- **Generative:** Agents that are defined with natural language using LLMs to give a real conversational feel to your chatbot.

Experiment with the Gemini API through **Google AI Studio** and **Vertex AI Studio.**

Reasoning loop: Prompt engineering techniques

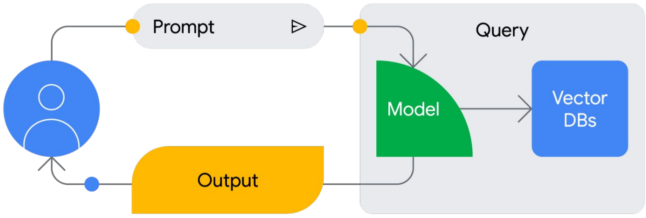
- **ReAct (Reason and act):** Allow the LLM to reason and take action on a user query.
- **CoT (Chain-of-thought):** Guide an LLM through a problem-solving process by providing examples with intermediate reasoning steps.
- **Metaprompting:** Use prompting to guide the AI model to generate, modify, or interpret other prompts.

Tooling

- **Extensions:** Connect to external services (via APIs).
- **Functions:** Define specific actions or tasks.
- **Data stores:** Provide access to information.
- **Plugins:** Add new skills and integrations

RAG: Retrieval-Augmented Generation

1. **Retrieval:** The LLM retrieves relevant information from external sources using tooling.
2. **Augmentation:** The retrieved information is incorporated into the prompt to the LLM.
3. **Generation:** The LLM processes the prompt and generates a response.
4. **Iteration (optional):** The LLM can iterate on the retrieval process as necessary.



Customer Engagement Suite: Tools to support your company in engaging with customers effectively, which can be built on top of Google's Contact Center as a Service (CCaaS), an enterprise-grade contact center solution that is native to the cloud.

- Conversational Agents: act as effective chatbots to your customers.
- Agent Assist: Support your live human contact center agents.
- Conversational Insights: Gain insights into all your communications with customers.

Vertex AI Search: Search and recommendation solutions for your business.

Gemini Enterprise: Integrate customized search and conversation agents that can access and understand data from various internal sources into your organization's internal websites or dashboards.

To **plan for your gen AI strategy**, establish a clear vision, prioritize use cases, invest in capabilities, manage change, measure value, and champion responsible AI.

Google Cloud

All Readings:
Introduction to Generative AI

Kindly note that the 30 minutes indicated on the platform considers the time that it may take you to browse through the reading resources provided. The total time required depends on the readings you decide to explore further.

Assembled readings on generative AI:

- Ask a Techspert: What is generative AI?
  https://blog.google/inside-google/googlers/ask-a-techspert/what-is-generative-ai/
- What is generative AI?
  https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai
- Google Research, 2022 & beyond: Generative models:
  https://ai.googleblog.com/2023/01/google-research-2022-beyond-language.html#GenerativeModels
- Building the most open and innovative AI ecosystem:
  https://cloud.google.com/blog/products/ai-machine-learning/building-an-open-generative-ai-partner-ecosystem
- Generative AI is here. Who Should Control It?
  https://www.nytimes.com/2022/10/21/podcasts/hard-fork-generative-artificial-intelligence.html
- Stanford U & Google's Generative Agents Produce Believable Proxies of Human Behaviors:
  https://syncedreview.com/2023/04/12/stanford-u-googles-generative-agents-produce-believable-proxies-of-human-behaviours/
- Generative AI: Perspectives from Stanford HAI:
  https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives
- Generative AI at Work:
  https://www.nber.org/system/files/working_papers/w31161/w31161.pdf
- The future of generative AI is niche, not generalized:
  https://www.technologyreview.com/2023/04/27/1072102/the-future-of-generative-ai-is-niche-not-generalized/

- The implications of Generative AI for businesses: https://www2.deloitte.com/us/en/pages/consulting/articles/generative-artificial-intelligence.html
- Proactive Risk Management in Generative AI: https://www2.deloitte.com/us/en/pages/consulting/articles/responsible-use-of-generative-ai.html
- How Generative AI Is Changing Creative Work: https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work

Assembled readings on large language models:

- NLP's ImageNet moment has arrived: https://thegradient.pub/nlp-imagenet/
- LaMDA: our breakthrough conversation technology: https://blog.google/technology/ai/lamda/
- Language Models are Few-Shot Learners: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
- Introducing Gemini: our largest and most capable AI model: https://blog.google/technology/ai/google-gemini-ai/#sundar-note
- The Power of Scale for Parameter-Efficient Prompt Tuning: https://arxiv.org/pdf/2104.08691.pdf
- Google Research, 2022 & beyond: Language models: https://ai.googleblog.com/2023/01/google-research-2022-beyond-language.html#LanguageModels
- Solving a machine-learning mystery: https://news.mit.edu/2023/large-language-models-in-context-learning-0207

Additional Resources:

- Attention is All You Need: https://research.google/pubs/pub46201/
- Transformer: A Novel Neural Network Architecture for Language Understanding: https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
- Transformer on Wikipedia:

https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)#:~:text=Transformers%20were%20introduced%20in%202017,allowing%20training%20on%20larger%20datasets.

- What is Temperature in NLP? https://lukesalamone.github.io/posts/what-is-temperature/
- Model Garden: https://cloud.google.com/model-garden
- Auto-generated Summaries in Google Docs: https://ai.googleblog.com/2022/03/auto-generated-summaries-in-google-docs.html